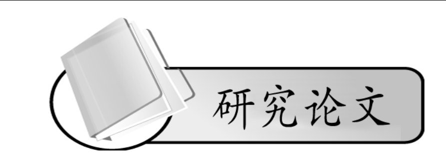


基于属性特征的评论文本情感极性量化分析*

李　慧　柴亚青

(西安电子科技大学经济与管理学院　西安 710126)

摘要: 【目的】从评论对象的属性特征出发解决情感极性量化问题。【方法】将在线评论文本分解构建三层评论体系, 即评论对象–对象属性–评论描述, 从属性层级抽取属性词集和对应的评论集, 考虑评论对象属性特征的不同影响, 引入属性因子, 并对 TFIDF 进行改进用以计算属性因子; 结合评论模式和评论语境提出基于属性特征的评论情感量化分析算法并采用 Python 语言予以实现。【结果】相较于传统机器学习分类算法(NB、SVM)、属性因子设置为等权重时, 本文算法在评论文本情感分类准确性方面有显著提高。【局限】评论集领域选择方面具有局限性, 量化算法在系数设定方面存在主观性。【结论】本文算法能有效解决情感极性量化问题, 进一步提高了情感分类准确性。

关键词: 评论文本　属性因子　评论模式　情感极性

分类号: G250

DOI: 10.11925/infotech.2096-3467.2017.0338

1　引　言

各类电商平台和社交网站每天都会产生大量的在线评论。通过对在线评论文本进行情感分析不仅能够辅助商家进行决策制定和网络营销, 还有助于舆情分析和帮助用户制定购买决策。情感分析是对在线评论文本进行研究分析的一个热点方向。早期的研究主要侧重于篇章级和句子级的整体情感判定, 但是不同用户在制定购买决策或选择服务时关注的属性特征并不相同。如果用户对其关注的产品属性特征的情感极性没有了解清楚前, 基本不会产生购买意向。对于商家来说, 商品或服务属性特征的优劣尤为重要, 商家如果没有从用户反馈中了解到其对产品或服务实质特征的情感倾向, 在制定产品或服务的改进方案时就不能够有针对性地制定方案, 反之也影响着商家绩效[1]。因此很多学者从属性特征角度研究在线评论, 识别评论文本中的属性特征词和情感词, 从而确定情感倾向。

当前针对属性特征的评论文本情感研究可以分为三类: 一是属性特征词的抽取算法研究; 二是从属性词角度出发, 通过构建领域情感词典进行评论挖掘分析; 三是从<属性, 情感词>对出发, 结合属性影响和语义进行情感倾向分析。在属性特征词抽取方面, Hu 等[2]利用规则提取出高词频的名词和名词性短语作为高频属性, 该方法的问题是属性词过于分散, 且没有进行归类筛选, 导致实验的准确度比较低。Ma 等[3]结合 LDA 与同义词林, 从 17 049 条数码相机评论中抽取属性, 以名词和名词性短语作为候选属性词, 采用 LDA 生成候选属性词列表, 结合同义词林对其进行扩展, 但是忽略了属性词上下文信息。周清清等[4]利用高频名词构建候选属性词, 通过深度学习构建候选属性

通讯作者: 柴亚青, ORCID: 0000-0001-8836-5124, E-mail: 1072386597@qq.com。

*本文系国家自然科学基金项目“基于可信语义 Wiki 的知识库构建方法与研究应用”(项目编号: 71203173)、中央高校基本科研业务费专项资金资助项目“大数据环境下基于主题模型的信息服务研究”(项目编号: JB160606)和国家自然科学青年基金项目“大规模动态社交网络社团检测算法研究”(项目编号: 71401130)的研究成果之一。

词向量，根据属性词向量完成候选属性词聚类，得到目标候选属性词集，这种方法能够更加全面发现评论对象细粒度属性，但在噪音过滤方面仍需加强，对于冷门属性效果较差。

由于属性词通常和情感词协同出现，且存在语义依存关系，娄德成等[5]运用依存关系对抽取属性词和情感词，采用手工构建属性特征层级结构。但手工构建方法耗时耗力，而且可移植性较差，倘若产品出现新的功能属性，需要调整原属性层级结构。有一类基于共现和极性传播的方法不仅考虑情感词之间的共现，还考虑情感词与情感对象之间的共现，认为在产品或服务的评论数据中，情感词和评论对象不会孤立出现，因此将情感词和评论对象进行协同抽取[6-7]。江腾蛟等[8]提出基于浅层语义与语法分析相结合的评价对象–情感词对抽取方法，设计语义角色标注和依存句法分析相结合的评价对象–情感词对抽取规则，在一定程度上解决了评价对象构成复杂性问题。在情感分类研究方面，不少学者考虑属性特征对情感倾向的影响，靳亚辉[9]在基于属性特征的产品评论挖掘研究时，设计基于属性的情感倾向确定算法，考虑情感词、程度词和否定词等，对于情感极性如何量化未涉及，也未考虑属性特征的影响差异。Parkhe 等[10]基于影评进行情感分析时引入驱动因子，利用抽取的特征词和情感词构建领域相关的特征–情感词表[11]，驱动因子高的评论特征对影评情感极性影响也越大。但该方法中的驱动因子值是实验中随机分配的，导致驱动因子的影响会随着实验不同而发生变化。王伟等[12]利用情感分析技术识别情感特征极性及其强度，结合产品特征的信息增益，建立产品特征评价对用户购买意愿的计量经济模型，得到产品属性特征重要度的量化方法。

上述研究主要侧重于属性特征的情感倾向判定，包括属性词的抽取，属性情感词对构建，属性特征影响的定性分析，基于属性特征的情感倾向判定等，存在两点不足：在情感量化分析方面未考虑不同属性特征对评论对象而言重要性存在差异；评论对象的情感倾向并非直接关联到情感词，通常是通过评论对象不同方面的情感描述组成。基于以上两点不足，本文从属性特征词抽取出发，结合评论分层思想和属性特征重要性差异，设计本文研究思路。

2 研究框架和思路设计

运用分层思想将在线评论分解为三层：评论对象；评论对象的不同方面(本文定为属性特征)；基于属性特征的情感表达[13]。从属性特征层级提出评论文本情感极性量化分析方法如下：

(1) 对评论集进行预处理(包括分词、去除停用词、词性标注、词频统计等)；

(2) 进行主观评论句筛选、属性特征词集和<Feature, Opinion>对抽取。考虑到评论对象的属性特征重要性差异，引入属性权重影响因子，简称为属性因子，并通过对 TFIDF 公式进行改进计算出属性因子值(改进依据：如果评论集中包含属性类中的属性词的评论数越多，该属性类越重要)，从而避免了实验中随机分配属性因子而产生的随机影响；

(3) 结合语义模型、属性因子、抽取出的<Feature, Opinion>集、评论语境、情感程度词、否定词、连词设计了情感极性量化算法；

(4) 采用标注好的评论语料库进行算法实现，计算出每条评论的情感量化分值并根据此分值确定其情感倾向，选择准确度(Accuracy)和 F 值(F-score)评价指标评价本文提出的情感极性量化算法，具体研究流程如图 1 所示。

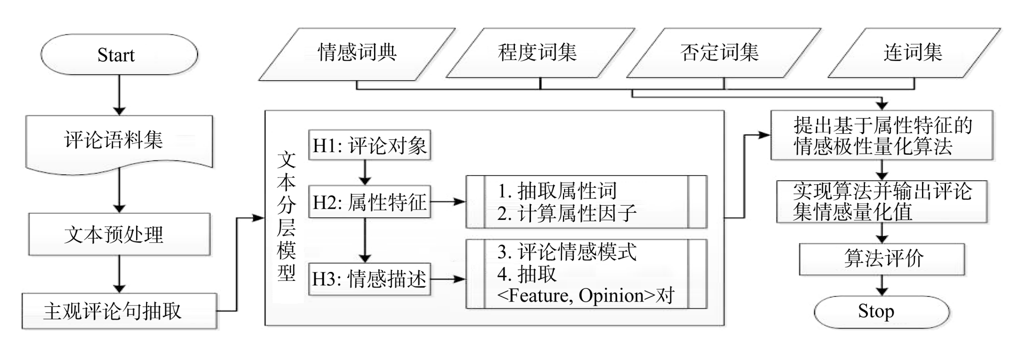


图 1　基于分层模型的评论挖掘流程

本文从属性特征层和情感描述层展开深入分析研究，具体研究可分为两部分：评论对象属性特征抽取和语境分析；提出基于属性特征的情感极性量化算法。

3 评论对象属性特征抽取和语境分析

基于评论对象的情感表达主要体现在产品属性的情感描述上，对于不同评论对象相同情感词可能表达出完全不同的情感，对于同一评论对象的不同属性特征，相同的情感词也可能产生截然相反的情感倾向。传统的情感分析方法直接基于情感词典进行分析，即筛选出评论中所有情感词并判断情感倾向，判断结果即为评论对象情感倾向，为了从细粒度角度提高情感分类准确性，本文运用三层模型构建评论体系[13]，如图 2 所示，即情感词只与属性特征相关，评论对象的情感与属性特征的情感相关。

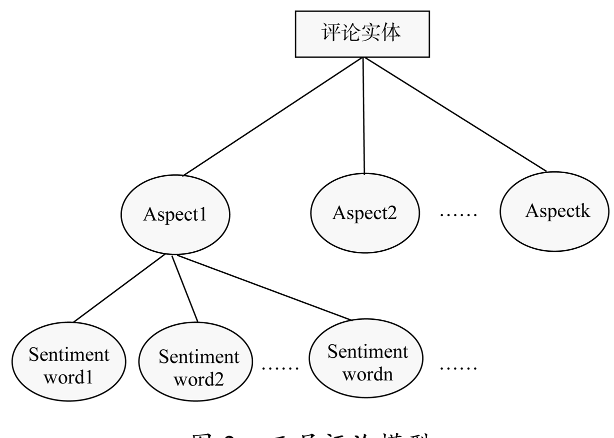


图 2 三层评论模型

3.1 评论对象属性特征

属性是对一个对象抽象的刻画，产品属性是产品性质的集合，通常从评论对象本质属性和非本质属性两个方面确定属性特征集 Z。本质属性是产品或服务区别于其他事物的属性；产品或服务的非本质属性包含多方面利益相关群体因素，如价格、销量、口碑等。一个评论对象的属性特征词集表示为：$Z=\{F_j\}_{j=1}^{K}$，其中 F_j 表示第 j 个属性特征词子集，j 的取值范围是 $[1,K]$，即将 Z 划分为 K 类；每个属性类又包含若干个子属性词，可表示为 $F_j=\{F_{j_1},F_{j_2},\cdots,F_{j_m}\}$，属性类层级结构表示如图 3 所示。

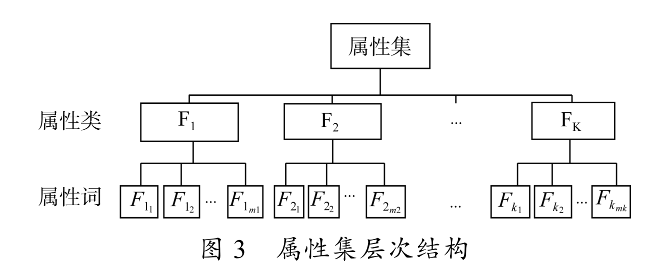


图 3 属性集层次结构

假定 $F_1,F_2,\cdots,F_K$ 是按照属性重要性排列，为了提高情感极性量化结果的准确性，引入属性类重要性因子 α，属性因子集表示为 $\{\alpha_j\}_{j=1}^{K}$，属性因子越大，属性特征的情感极性和情感极性值对评论集的情感分类结果影响力越大。

(1) 确定属性特征词集

根据网站汇总的用户点评信息和产品详情确定属性类和初步属性集，例如：携程旅行网中某酒店的用户点评信息汇总结果为：位置(4.4)、设施(4.5)、服务(4.4)、环境(4.6)(括号内分别为各项得分)，酒店详情包含：服务、设施、交通等项目概述，可将该酒店属性类确定为：{位置、设施、服务、环境、交通}；通过统计词频筛选出评论集中的所有名词和名词性短语作为候选词集，通过点互信息(PMI)识别出与评论实体互信息值高的名词和名词性短语作为候选产品属性特征词集[14]，将搜狗细胞词库下载的该领域最新词汇加入该候选集予以扩充，结合人工判定、同义词林等将扩充后的候选属性词再次进行筛选并加入到属性词集中。点互信息计算如公式(1)所示。

$$PMI(F_j,ph)=\log\frac{p(F_j,ph)}{p(F_j)\times p(ph)} \quad j=1,2,\cdots,K \quad (1)$$

其中，ph 为评论集中的名词和名词性短语，$p(F_j,ph)$ 为评论集中属性类 F_j 和名词或名词性短语 ph 共同出现的概率，$p(F_j)$ 为属性类 F_j 出现的概率，$p(ph)$ 为候选属性词出现的概率。每条评论中候选词出现一次或多次均记为一次。

(2) 计算属性因子

由于在线评论都为短文本，传统长文本的特征权重方法不再适用。本文设计属性因子计算方法采用的依据为：如果属性词词频越大，且包含该属性词的文档数越多，则该属性词重要程度越高。该依据基于假设：针对一个特征项，它在一个文档中出现很多次，同时也出现在多个文档中，那么该特征词具备较大的

区分度[15]。

TFIDF 是一种衡量特征项权值的有效方法，特征项的重要性随着它在文档中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，TF 表示特征项词频，为了抑制噪声加权，引入 IDF 表示逆文档频率[15]。而属性因子权重随着属性词频数和文档频数均成正比增加，即：如果属性类子属性词词频和越高，且评论集中包含属性类子属性词的评论数越多，该属性类则越重要。因此用 TF 表示文档频率，同理为了抑制噪声加权，采用 $IDF^{(-1)}$ 表示正文档频率(随着文档频率的增加而增大)，同时为了避免属性因子差异太大，引入归一化思想，使得 $\sum_{j=1}^{K}\alpha_j=1$，α_j 计算如公式(2)和公式(3)所示。

$$\alpha_j=\frac{tf_{F_{j,i}}\times(idf_j)^{(-1)}}{\sum_{j=1}^{K}tf_{F_{j,i}}\times(idf_j)^{(-1)}} \tag{2}$$

$$idf_j=\log\frac{N}{\left|\{i:F_j\in C_i\}\right|+0.5} \tag{3}$$

其中，$tf_{F_{j,i}}$ 表示 F_j 在评论集 $\{C_i\}_{i=1}^{N}$ 中的频率(即 F_j 子属性词词频之和)，N 表示评论集的总数，$\{i:F_j\in C_i\}$ 表示 $\{C_i\}_{i=1}^{N}$ 包含 F_j 的评论数。

3.2 主观评论中<Feature, Opinion>集抽取

在线评论文本不仅包含主观评论句，也包括中立性的客观性语句，还有一些不相关干扰性语句。因此首先需要剔除客观评论和不相关评论，然后对评论文本进行情感词抽取。假定一个句子中出现情感词(褒义或贬义)，它就是主观评论句，反之，则为客观或不相干评论句。本文依据情感词集中的情感词作为评判依据，抽取出评论集中的所有主观评论。

获取主观评论文本情感词(评论词)的方法主要分为两种：一种是直接抽取主观评论中的所有情感词；另一种是先确定评论对象，然后有针对性地抽取关于评论对象的情感词[16]。本文采取有针对性的情感词抽取方法，结合评论情感模式进行情感词抽取，评论情感模式参照 Turney[17]提出的双词模式，如表 1 所示。

其中，JJ 表示形容词，NN 表示名词，NNS 表示名词复数，VB 表示动词，VBD、VBN、VBG 分别表示动词过去时、过去分词和动名词，RB 表示形容词，RBR 和 RBS 分别表示副词比较级和副词最高级。具体抽取过程为：根据确定好的属性特征词集，定位到所有包含属性词的主观评论句，根据表 1 的双词情感模式抽取出所有基于属性特征的情感评论。抽取规则为：属性词 F_{j_m} 所处评论句包含 L 个词，属性词的所处位置可表示为：$\{w_1,w_2,\cdots,w_l,F_{j_m},w_{l+2},w_{l+3},\cdots,w_L\}$，针对 F_{j_m} 及其前后词语进行模式匹配，分别进行前向匹配和后向匹配，符合表 1 中任一种模式则抽取出该条评论句中 $\{w_{l-1},w_l,F_{j_m},w_{l+2},w_{l+3}\}$ 作为属性词的情感评论，F_{j_m} 和抽取的 $\{w_{l-1},w_l,F_{j_m},w_{l+2},w_{l+3}\}$ 构成 <Feature, Opinion>集。

表 1　双词情感模式

	第一个词	第二个词	第三个词
模式 1	JJ	NN, NNS	anything
模式 2	RB, RBR, or RBS	JJ	not NN or NNS
模式 3	NN or NNS	JJ	not NN or NNS
模式 4	JJ	JJ	not NN or NNS
模式 5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

具体匹配过程举例：

首先将 $\{w_1,w_2,\cdots,w_l,F_{j_m},w_{l+2},w_{l+3},\cdots,w_L\}$ 与模式 1 进行匹配，若 w_l 为形容词，则匹配成功，抽取 $\{w_{l-1},w_l,F_{j_m},w_{l+2},w_{l+3}\}$；若不满足模式 1，则将 $\{w_1,w_2,\cdots,w_l,F_{j_m},w_{l+2},w_{l+3},\cdots,w_L\}$ 与模式 2 进行匹配，此时从 F_{j_m} 开始进行后向匹配，若 $w_{l+2}w_{l+3}$ 分别为副词和形容词，则匹配成功；若不满足模式 2，则继续与模式 3 进行匹配，此时 F_{j_m} 为第一个词，若 w_{l+2} 为形容词，则匹配成功，否则与模式 4 进行匹配，此时 w_{l+2} 为第一个词，若 $w_{l+2}w_{l+3}$ 均为形容词，则匹配成功，否则与模式 5 进行匹配，可进行前向匹配或后向匹配，如果 $w_{l-1}w_l$ 或 $w_{l+2}w_{l+3}$ 分别为副词和动词，均能匹配成功。

3.3 基于连词的 POS 标注

在计算属性特征情感极值时还应该考虑属性特征词所在的评论语境，本文主要考虑语境中连词对情感的影响。主观评论句中如果存在连词(主要考虑转折连词、让步关系连词)，则会对评论句的情感倾向产生影响，转折连词会改变评论情感方向，递进连词会增强情感倾向[18]。例如：酒店服务很好，但是地理位置特

别差，此处转折连词"但是"使评论句情感倾向发生由正向到负向的转变且负向情感倾向增强。考虑到语境中不同连词对情感倾向的影响，为了提高情感极值计算的准确性，对主观评论文本中的连词进行标注，即先构建包含转折、递进关系的连词词库，如表 2 所示。

表 2 连词词汇表

类型	连词
转折	但是、偏偏、只是、不过、至于、不料、岂知、虽然、然而、而、即使、但、可是、不过、却
递进	而且、更、更加、并、甚至、不如、不及、乃至、并且、况、况且、何况

根据连词标注规则进行匹配，具体的标注规则如下：

①转折连词：如果连词英文为"but"，中文为"但是"、"偏偏"、"岂知"时，连词后面表达的情感是评论者侧重的情感倾向。即：前肯定后否定极性为否定，前否定后肯定极性为肯定；

②递进连词：如果连词英文为"even、also、in addition"等，中文为"而且、更、甚至"等类似词时，连词后面表达的情感是评论者侧重的情感倾向。

4 评论文本情感极性量化算法

4.1 基于属性特征的情感极性量化

评论中如果存在某个属性的描述，则包含该属性 $F_j=\{F_{j_1},F_{j_2},\cdots,F_{j_m}\}$ 中一个或多个属性特征词。假定一条评论中如果存在某个属性的属性特征词的情感表达，则该属性特征词的情感倾向是确定且唯一的。将根据抽取规则抽取出<Feature, Opinion>集拆分为 $<F,S,P>$ 对，S 为情感词(评论词)，通常为形容词，P 为表达 S 程度的修饰词，一般为程度副词和否定词。程度副词是体现情感强度的重要指标，对情感词起到修饰作用，通常划分为强化修饰和弱化修饰，如<外观，美，很>和<外观，美，稍微>两个抽取结果，很美和稍微美的情感程度显然不同。单个否定词直接对情感词进行否定修饰，若出现双重否定则情感倾向不发生变化。本文将否定词与程度副词一并看成情感修饰词。

根据抽取<Feature, Opinion>集中的 $<F,S,P>$ 对计算每条评论的属性情感极性值，先计算 $<F,S>$ 情感极性，然后计算 $<S,P>$ 极性权重，将极性权重与属性因子相乘得到属性情感极性值。

(1) 确定属性特征情感极性

如公式(4)和公式(5)所示，${Pos_{F_j}}^i$ 表示第 i 条评论中属性 F_j 的情感为褒义的权重，${Neg_{F_j}}^i$ 表示第 i 条评论中属性 F_j 的情感为贬义的权重，${FPos_{F_j}}^i$ 表示第 i 条评论的 $<F_j,S,P>$ 对中的情感词 S 出现在褒义词集中的频率，${FNeg_{F_j}}^i$ 表示第 i 条评论的 $<F_j,S,P>$ 对中的情感词 S 出现在贬义词集中的频率。

$$
\begin{aligned}
{Pos_{F_j}}^i &= \frac{{FPos_{F_j}}^i}{{FPos_{F_j}}^i+{FNeg_{F_j}}^i}, \\
{Neg_{F_j}}^i &= \frac{{FNeg_{F_j}}^i}{{FPos_{F_j}}^i+{FNeg_{F_j}}^i}
\end{aligned} \tag{4}
$$

$$
{S_{F_j}}^i = {Pos_{F_j}}^i - {Neg_{F_j}}^i \tag{5}
$$

${FPos_{F_j}}^i$ 和 ${FNeg_{F_j}}^i$ 具体计算方法如下：由于第 i 条评论中可能出现属性类 F_j 的多个子属性词，且其所对应的情感词褒贬义各异，因此分别计算出其对应情感词出现在褒、贬义词集中的频率之和即为 ${FPos_{F_j}}^i$ 和 ${FNeg_{F_j}}^i$。

根据 $sign({S_{F_j}}^i)$ 函数确定第 i 条评论中属性 F_j 的情感极性，1 表示情感极性为正，0 表示情感极性为中性，–1 表示情感极性为负，如公式(6)所示。

$$
sign({S_{F_j}}^i)=\begin{cases} 1 & S_{F_j}>0 \\ 0 & S_{F_j}=0 \\ -1 & S_{F_j}<0 \end{cases} \tag{6}
$$

(2) 情感极性值计算

评论 c_i 的所有 $<F_j,S,P>$ 对中情感修饰词 P 的权重用 W 表示，若修饰词为程度副词，则根据程度系数确定程度副词权重，蔺璜等[19]提出把程度副词划分为 6 个等级，本文在该划分基础上为每个等级程度副词定义了权重即程度系数，以便情感程度量化计算；若修饰词为否定词，则根据否定词个数确定修饰词权重。考虑到语境问题，假定每条评论中不同属性的情感相互独立，本文针对目标语句(即：包含属性特征的评论语句)内的连词类别设置连接系数，不再考虑目标语句外的连词影响。根据 3.3 节中构建的连词库设定连词影响程度系数，转折连词表示情感倾向相反且情感程度加深，递进关系连词表示情感倾向增强，刘玉娇等[20]在基于情感词典与连词结合进行中文文本情感分类时，考虑转折连词和递进连词对情

感值判定的影响，将转折连词情感影响程度设定为–1，递进连词情感影响程度设定为 1.5，本文采用该设定值对连词系数进行设定。连词的连接系数确定如公式(7)所示。

$$W_{conj}{}^{i}(F_j)=\begin{cases}-1 & 连词为转折连词 \\ 1.5 & 连词为递进连词 \\ 1 & 其他\end{cases} \quad (7)$$

评论 c_i 基于 F_j 的情感极性量化公式如公式(8)所示。

$$\overline{W_{F_j}^{i}}=\frac{\sum_{m=1}^{n(F_j)} W(F_j,m)\times SO(F_j,m)\times W_{conj}(F_j,m)}{n(F_j)} \quad (8)$$

其中，鉴于一条评论句中可能出现某个属性类中一个或多个属性词的评论，令 $n(F_j)$ 为属性类 F_j 在评论句 c_i 中出现的总次数，$W(F_j,m)$ 为评论句中第 m 次出现属性类 F_j 中的属性词时所对应的情感修饰词的权重，$SO(F_j,m)$ 为评论中第 m 次出现属性类 F_j 中的属性词时所对应的情感值，即情感词所对应的褒义或贬义权重，$W_{conj}(F_j,m)$ 为评论中第 m 次出现属性类 F_j 中的属性词时，其所在评论语境中(转折或递进)连词的权重。$\overline{W_{F_j}^{i}}$ 为第 i 条评论基于属性类 F_j 所对应的平均情感极性值。

评论语料集 $\{c_1,c_2,\cdots,c_n\}$ 中每条评论 c_i 基于属性集 $\{F_1,F_2,\cdots,F_K\}$ 的情感极性值计算如公式(9)所示。

$$Score_{c_i}=\sum_{j=1}^{K}\alpha_j\times\overline{W_{F_j}{}^{i}} \quad (9)$$

4.2 基于属性特征的情感极性量化算法描述

输入：评论语料集 $\{c_1,c_2,\cdots,c_n\}$，属性集 $\{F_1,F_2,\cdots,F_K\}$，语料集中每条评论中的 $<Feature,Opinion>$ 集，情感词典，程度副词集，否定词集，连词集；

输出：$<C, Score>$ 情感倾向程度量化数据库。

Begin：$i=1,\cdots,N$；$j=1,2,\cdots,K$；F_j 表示第 j 个属性；

①计算属性因子 α_j，采用改进的 TFIDF，令 i=1, j=1;

②扫描属性集 $\{F_1,F_2,\cdots,F_K\}$：根据情感词典判定 $<Feature,Opinion>$ 中 F_j 的褒贬义，并初始化 $score(F_j)$，初始化规则为：F_j 对应的 $Opinion$ 情感词为褒义词，初始化 $score(F_j)=1$；F_j 对应的 $Opinion$ 情感词为贬义词，初始化 $score(F_j)=-1$；否则 $score(F_j)=0$；

③扫描程度副词集：更新 $score(F_j)$，更新规则为：$score(F_j)=r\times score(F_j)$，$r$ 为程度副词的程度系数；

④扫描否定词集：更新 $score(F_j)$，更新规则为：$score(F_j)=(-1)^n\times score(F_j)$，$n$ 为否定词个数；

⑤扫描连词库：更新 $score(F_j)$，更新规则为：如果连词为转折连词，更新 $score(F_j)=-1\times score(F_j)$；如果连词为递进连词，更新 $score(F_j)=1.5\times score(F_j)$；

⑥如果 j=K，计算 $\overline{score(F_j)}$ 如公式(10)所示，然后执行步骤⑦；如果 j<K, j=j+1，重复步骤②–步骤⑤；

$$\overline{score(F_j)}=\frac{\sum_{m=1}^{n(F_j)} score(F_j,m)}{n(F_j)} \quad (10)$$

其中，$\overline{score(F_j)}$ 表示第 i 条评论基于属性 F_j 的情感量化评论分值。

⑦计算第 i 条评论的情感倾向程度值 $score_i$，如公式(11)所示。

$$score_i=\sum_{j}\alpha_j\times\overline{score(F_j)} \quad (11)$$

⑧如果 $i=N$，则 stop，输出 $\{score_i\}$；否则：$i=i+1$，继续扫描语料库的 $<Feature,Opinion>$ 词对集，重复步骤②–步骤⑦。

5 实验验证与结果分析

5.1 数据集选择

本文选择中国科学院大学谭松波教授发布的数据集，内容为酒店评论语料，可从数据堂网站下载，该数据集由谭松波老师收集整理标注，语料来自携程网，自动采集并经过整理而成，语料集规模为 10 000 篇，为了方便研究，语料集被整理成 4 个子集，子集记录之间有重复，前三个都是平衡数据集，第 4 个是非平衡数据集，如表 3 所示。正面评论示例："房间内环境还是不错的，就是上网有点贵，12 块一个小时，还有门口修路，门前环境不好。"；负面评论示例："房间装修陈旧，下水管堵塞，晚上折腾了 2 个多小时，还是没有修好。"。涉及的属性特征有"内外环境""网络""装修""设施"，例如："上网有点贵"中"贵"为情感词，"有点"为程度副词。

表 3 评论语料集

数据集	标注好的正面评论数目	标注好的负面评论数目
chnSenticorp2000	1 000	1 000
chnSenticorp4000	2 000	2 000
chnSenticorp6000	3 000	3 000
chnSenticorp10000	3 000	7 000

对评论语料库进行文本预处理(包括分词、去除停用词、词频统计、词性标注等), 采用 3.1 节中的属性特征抽取方法, 选择携程旅行官网的评论汇总信息和酒店详情描述, 确定出评论对象的 7 个属性特征类为: 环境、设施、餐饮、交通、服务、价格、位置, 将词频统计结果中的高频名词和名词性短语进行点互信息(PMI)计算, 结合同义词林和酒店特征, 抽取出 7 个属性类中的子属性词, 再结合搜狗细胞库下载的酒店专用名词对属性词集进行扩充, 结果如表 4 所示。

表 4 本文抽取的属性词集

属性(Feature)	属性词
F1: 环境	风景、环境、氛围、外观、外表、条件、卫生、空气、酒店环境、酒店氛围、宾馆、周围、周围环境、周边环境、大堂、大堂环境、外观、门面、室内环境、室内、屋内、房子、房间、楼道、走廊、气味、味道、霉味、油漆味、烟味、噪音、噪声
F2: 设施	设施、设计、风格、配套、设备、设置、布置、装置、配备、装备、内饰、内里、建筑、格局、硬件、硬件设施、软件、软件设施、装修、卧具、家具、电梯、客房、标准间、房间面积、房间大小、光线、空间、电视、网络、网速、上网、宽带、空调、墙壁、墙纸、床、毛巾、床单、被罩、被褥、地毯、地板、地面、卫生间、洗手间、厕所、浴室、淋浴、浴缸、热水、洗澡、洗漱用品、个人用品、房间隔音、隔音、停车场、停车、周围设施、通风
F3: 餐饮	餐饮、就餐、餐厅、饭菜、上菜、点餐、叫餐、早餐、早茶、早点、早饭、自助餐、下午茶、饮食、味道、品种、种类、吃饭
F4: 交通	交通、周围交通、路线、出行、外出、打车、进出、购物、景点
F5: 服务	服务态度、态度、表情、语气、口气、服务意识、服务员态度、服务、服务水平、素质、服务素质、前台、服务员、门童、服务生、前台服务、酒店服务、管理、退房、客服
F6: 价格	价格、收费、价钱、价位、性价比、房价、结账、账单、手续
F7: 位置	地理位置、位置、地位、地点、地方、地段、场所、火车站、机场

情感修饰词系数设定方法如 3.1 节所述, 将知网(HowNet)的 219 个程度副词和评论集中筛选出的程度副词结合构成程度副词集划分为 6 个等级, 程度系数依次设置为: 2、1.5、1.25、1.2、0.8、0.5, 若评论中不含程度副词, 则令程度系数为 1, 否定词程度系数统一设定为–1。连词连接系数设定见 4.1 节。情感词典选择知网和 NTUSD 情感词, 如表 5 所示。

表 5 情感词典

情感词典	积极词汇	消极词汇	总数
HowNet	4 566	4 370	8 851
NTUSD	2 846	8 325	10 027

评价指标: 实验结果评价指标选择信息检索领域传统的 Accuracy 和 F_1 值对实验结果进行比较分析, 本文选择的评价指标含义和信息检索中的具体含义一致, 根据表 6 中的混淆矩阵计算 Accuracy 和 F_1 值[21], 计算如公式(12)至公式(15)所示。

表 6 混淆矩阵

	Correct label	
	True	False
Positive	TP(True Positive)	FP(False Positive)
Negative	TN(True Negative)	FN(False Negative)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Recall = \frac{TP}{TP + TN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

其中, TP、TN 分别表示预测正确的正向类别和负向类别数; FP、FN 分别表示预测错误的正向类别数和负向类别数。

5.2 实验结果及分析

(1) 本文量化算法实现结果

基于标注好的酒店评论语料库进行训练，首先进行文本预处理，采用 Python 语言进行分词、去除停用词、词频统计、词性标注，根据抽取的属性词集，提取出评论集的所有 $<\mathit{Feature},\mathit{Opinion}>$ 对；然后依据选择的情感词典、程度副词词典、否定词集、连词集，将本文提出的情感强度量化算法用 R 语言进行实现，计算出评论集中每条评论基于属性特征的情感极性值。并根据计算结果判定每条评论的情感倾向；最后结合原始已标注的语料库进行对比研究，情感分类结果通过计算混淆矩阵的 Accuracy 和 F_1 指标予以评价。

为了验证本文提出情感极性量化算法的实现效果和属性因子对算法结果的影响，设置两组对照实验：

①对照实验 1：改进本文算法中的属性因子设置方法，将属性因子设为等权重，其余过程不发生变化；

②对照实验 2：采用有监督机器学习(主要用 SVM 和 NB)方法分别对语料库进行多次情感分类训练。

本文参考文献[15]的中文网络评论的情感倾向分析研究中的训练模型训练语料库得出分类结果，计算混淆矩阵的 Accuracy 和 F_1 值并评价本文情感量化算法。实验结果如表 7–表 10 所示。

表 7 属性因子计算结果

属性	环境	设施	餐饮	交通	服务	价格	位置
属性因子	0.501406	0.042195	0.029424	0.005845	0.389272	0.019860	0.011962
属性因子(对照)	0.142857	0.142857	0.142857	0.142857	0.142857	0.142857	0.142857

表 8 部分情感极值量化示例

评论序列	预处理后的评论	提取属性情感对	POS 标注	计算情感极值	情感分类	备注
Comment1	|风景还算不错|酒店早餐很难吃	<风景，不错，还算> <早餐，难吃，很>	无	–0.305203935	N	1 表示无程度副词
Comment2	|房间家具太差|早餐质量太差|环境好但交通太差	<家具，大，1> <早餐，差，太> <环境，好，1> <交通，差，太>	但：转折连词	–1.532515171	N	无
Comment3	|但房间里的淋浴设施不好|前台小姐服务很不好|服务意识太差	<设施，不好，1> <服务，不好，很> <服务意思，差，太>	无	–2.035849256	N	无
Comment4	|环境比较温馨|房间比较干净|卫生间设施较完备	<环境，温馨，比较> <房间，干净，比较> <设施，完善，较>	无	0.709084635	P	无
Comment5	|虽然房间的条件略显简陋|但环境、服务还有饭菜都还是很不错的	<条件，简陋，1> <环境，不错，很> <服务，不错，很> <饭菜，不错，很>	但：转者连词	0.405485228	P	无

表 9 实验对比结果(Accuracy)

语料库	Accuracy			
	属性因子等权重	传统分类方法 NB	传统分类方法 SVM	本文算法
chnSenticorp2000	88.33%	0.791	0.879	89.23%
chnSenticorp4000	89.56%	0.832	0.881	89.90%
chnSenticorp6000	90.01%	0.854	0.908	91.45%
chnSenticorp10000	91.59%	0.873	0.911	92.88%

表 10 实验对比结果(F_1)

语料库	F_1			
	属性因子等权重	传统分类方法 NB	传统分类方法 SVM	本文算法
chnSenticorp2000	80.32%	0.732	0.793	81.13%
chnSenticorp4000	80.57%	0.792	0.801	82.60%
chnSenticorp6000	82.31%	0.801	0.818	84.25%
chnSenticorp10000	82.69%	0.809	0.821	85.19%

(2) 实验结果分析

分析表 7–表 10, 表 7 中除表头外第一行为采用改进的 TFIDF 权重计算方法计算出的各属性的权重因子, 第二行为设置的等权重属性因子值作为对照实验 1 与本文提出的 TFIDF 改进方法予以比较; 表 8 为本文提出的量化算法计算情感得分的部分示例; 表 9 和表 10 为本文提出的量化算法进行情感分类的结果与对照实验进行情感分析的结果对比。总体来看, 本文在选择 Accuracy 和 F_1 作为评价指标的情况下, 基于提出的基于属性特征的情感极性量化算法进行情感分类准确性明显高于传统机器学习方法, 本文采用改进的 TFIDF 计算属性因子, 将其用于量化算法中的分类准确率明显高于等权重属性因子用于量化算法中的分类准确率。

分析表 9 和表 10, 对照实验 1 与本文算法结果对比表明, 属性因子在情感极性量化计算时产生较大影响, 将属性因子设置为等权重时的情感分类准确率低于属性因子不等权重时的分类准确率, 这与评论对象的不同属性特征会对购买意向产生不同影响相吻合, 即: 对于产品或服务, 人们总是首先关注重要性程度高的属性特征评论, 然后关注重要性程度低的特征。本文选择酒店评论语料, 计算出的属性因子中环境、服务属性类的重要性程度比较高, 设施重要性次之, 而交通、位置、餐饮的重要性程度相对较低, 这也与社会发展趋势相吻合, 人们追求高质量服务、环境和设施高于低层面的需求。

分析表 9 和表 10, 对照实验 2 与本文算法结果表明, 本文提出的基于属性特征的情感量化算法在情感分类准确率方面高于传统分类方法, F_1 值相较于传统分类方法也相应提高。传统情感分析方法提取出的文本情感特征既包含属性情感特征又包含非属性情感特征, 而本文针对属性词集抽取出评论中<Feature, Opinion>对, 从细粒度角度研究, 消除了非属性情感特征的影响, 设计基于属性特征的情感量化算法, 算法实现结果表明本文提出的基于属性特征的情感极性量化算法在情感分类方面具有较高准确性。

6　总结及展望

本文从评论对象属性级别进行情感极性量化分析, 在更细的粒度上研究用户针对产品特征的评价。基于三层评论模型, 即情感词只与属性特征相关, 评论对象的情感与属性特征的情感相关的前提, 主要针对属性情感极性如何量化进行算法设计。从属性层次角度出发抽取出属性词集和基于属性词的评论集, 结合属性词的重要性程度、情感词、情感程度词、否定词, 结合语境加入连词, 在计算属性因子时采用改进的 TFIDF 权重计算方法, 设计基于属性特征的情感极性量化算法。最后在标注好的语料库上进行实验验证, 通过设置两组对照实验进行比较分析, 验证了本文提出的算法在评论文本情感分类方面取得了比较高的准确率。

下一步研究计划: 在语境方面, 考虑语句特殊句型, 如是否为反问句、感叹句等会对评论句情感产生的影响, 在研究属性特征词情感倾向时考虑如何将属性层级情感和评论文本情感倾向结合; 在领域情感词典扩充方面, 考虑不同领域特定情感词和普通情感词结合, 扩充领域情感词集; 属性特征词集抽取方面, 对现存的属性词抽取算法予以改进, 抽取出评论对象的更加完整的属性词集。

参考文献:

[1] 孟园, 王洪伟, 王伟. 网络口碑对产品销量的影响: 基于细粒度的情感分析方法[J]. 管理评论, 2017, 29(1): 144-154. (Meng Yuan, Wang Hongwei, Wang Wei. The Effect of Electronic Word-of-Mouth on Sales Through Fine-Gained Sentiment Analysis [J]. Management Review, 2017, 29(1): 144-154.)

[2] Hu M, Liu B. Mining and Summarizing Customer Reviews [C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA. 2004: 168-177.

[3] Ma B, Zhang D, Yan Z, et al. An LDA and Synonym Lexicon Based Approach to Product Feature Extraction from Online Consumer Product Reviews [J]. Journal of Electronic Commerce Research, 2013, 14(4): 304-314.

[4] 周清清, 章成志. 在线用户评论细粒度属性抽取[J]. 情报学报, 2017, 36(5): 484-493. (Zhou Qingqing, Zhang Chengzhi. Fined-Grained Aspect Extraction from Online Customer Reviews[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(5): 484-493.)

[5] 娄德成, 姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用, 2006, 26(11): 2622-2625. (Lou Decheng, Yao Tianfang. Semantic Polarity Analysis and Opinion on Chinese Review Sentences [J]. Computer

Applications, 2006, 26(11): 2622-2625.)

[6] Lazaridou A, Titov I, Sporleder C. A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 1630-1639.

[7] Xu L, Liu K, Lai S, et al. Walk and Learn: A Two-Stage Approach for Opinion Words and Opinion Targets Co-Extraction[C]// Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 95-96.

[8] 江腾蛟, 万常选, 刘德喜, 等. 基于语义分析的评价对象–情感词对抽取[J]. 计算机学报, 2017, 40(3): 617-633. (Jiang Tengjiao, Wan Changxuan, Liu Dexi, et al. Extracting Target-Opinion Pairs Based on Semantic Analysis[J]. Chinese Journal of Computers, 2017, 40(3): 617-633.)

[9] 靳亚辉. 基于属性集合的产品评论挖掘研究[D]. 武汉: 华中科技大学, 2011. (Jin Yahui. Product Review Mining Based on Feature Set[D]. Wuhan: Huazhong University of Science and Technolgy, 2011.)

[10] Parkhe V, Biswas B. Sentiment Analysis of Movie Reviews: Finding Most Important Movie Aspects Using Driving Factors[J]. Soft Computing, 2016, 20(9): 1-7.

[11] 王文华, 朱艳辉, 徐叶强, 等. 基于 SVM 的产品评论属性特征的情感倾向分析[J]. 湖南工业大学学报, 2012, 26(5): 76-80. (Wang Wenhua, Zhu Yanhui, Xu Yeqiang, et al. Analysis on Emotional Tendency of Attribute Characteristics in Product Reviews Based on SVM [J]. Journal of Hunan University of Technology, 2012, 26(5): 76-80.)

[12] 王伟, 王洪伟. 特征观点对购买意愿的影响: 在线评论的情感分析方法[J]. 系统工程理论与实践, 2016, 36(1): 63-76. (Wang Wei, Wang Hongwei. The Influence of Aspect-Based Opinions on User's Purchase Intention Using Sentiment Analysis of Online Reviews [J]. Systems Engineering——Theory & Practice, 2016, 36(1): 63-76.)

[13] Yang K, Cai Y, Huang D, et al. An Effective Hybrid Model for Opinion Mining and Sentiment Analysis[C]// Proceedings of the 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). 2017.

[14] Hu M, Liu B. Mining Opinion Features in Customer Reviews[C]//Proceedings of the 19th National Conference on Artifical Intelligence. 2004: 755-760.

[15] 陈贤. 中文网络评论的情感倾向性分析研究[D]. 北京: 北京邮电大学, 2014. (Chen Xian. Research on Chinese Online Reviews Sentiment Classification [D]. Beijing: Beijing University of Posts and Telecommunications, 2014.)

[16] 陆叶, 张晓如. 基于语义文法的属性知识获取[J]. 信息技术, 2017, 41(2): 38-42. (Lu Ye, Zhang Xiaoru. Acquiring Attributes Knowledge Based on Semantic Grammar [J]. Information Technology, 2017, 41(2): 38-42.)

[17] Turney P D. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 417-424.

[18] Quan C, Ren F. Feature-level Sentiment Analysis by Using Comparative Domain Corpora[J]. Enterprise Information Systems, 2016, 10(5): 505-522.

[19] 蔺璜, 郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报: 哲学社会科学版, 2003, 26(2): 71-74. (Lin Huang, Guo Shuhui. On the Characteristics Range and Classification of Degree[J]. Journal of Shanxi University: Philosophy & Science, 2003, 26(2): 71-74.)

[20] 刘玉娇, 琚生根, 伍少梅, 等. 基于情感字典与连词结合的中文文本情感分类[J]. 四川大学学报: 自然科学版, 2015, 52(1): 57-62. (Liu Yujiao, Ju Shenggen, Wu Shaomei, et al. Classification of Chinese Texts Sentiment Based on Semantic and Conjunction[J]. Journal of Sichuan University: Natural Science Edition, 2015, 52(1): 57-62.)

[21] Chen T, Xu R, He Y, et al. Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis[J]. IEEE Computational Intelligence Magazine, 2016, 11(3): 34-44.

作者贡献声明:

李慧: 提出研究思路, 修订论文;
柴亚青: 采集数据, 设计并实现算法, 分析实验结果, 起草论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 1072386597@qq.com。

[1] 李慧, 柴亚青. ChnSentiCorp.rar. 中文情感挖掘酒店评论语料.
[2] 李慧, 柴亚青. NTUSD.txt, 知网.txt. 情感词典.
[3] 李慧, 柴亚青. 连词.txt, 程度词.txt, 否定词.txt, 停用词.txt.
[4] 李慧, 柴亚青. 量化结果.xlsx. 评论集情感量化结果.

收稿日期: 2017-04-26
收修改稿日期: 2017-08-15

Analyzing Sentiment Polarity of Comments Based on Attributes

Li Hui　Chai Yaqing
(School of Economics and Management, Xidian University, Xi'an 710126, China)

Abstract: [**Objective**] This article tries to quantitatively study the sentiment polarity of online comments base on the targets' attributes. [**Methods**] First, we analyzed the comments by their objects, attributes and contents. Then, we extracted the attribute words and the corresponding comment sets. Third, we introduced the attribute factors and calculated their values with the modified TFIDF formula. Finally, we developed a quantitative analysis algorithm based on the attribute features with Python. [**Results**] Compared to the traditional machine learning classification algorithms (e.g., NB and SVM), our method improved the accuracy of sentiment classification, when the attribute factor was set to equal weight. [**Limitations**] The comments selection method and the coefficients parameters of the proposed algorithm need to be improved. [**Conclusions**] Our method could effectively improve the accuracy of the sentiment classification.

Keywords: Comment Text　Attribute Factor　Comment Mode　Sentiment Polarity

---

自出版图书在 2015 年-2016 年间上涨 8%

根据 ProQuest 子公司 Bowker 的最新报告，自 2011 年以来，自出版(Self-Publishing，是指作者在没有第三方出版商介入的情况下，利用电子图书平台自主出版书籍或多媒体产品，也称为"原生电子书")的国际标准书号(ISBN)的数量上涨了 218.33%。2016 年，共有 786 935 份 ISBN 号分配给自出版的作品；而在 2011 年，这个数字仅是 247 210。

Bowker 这项新的研究凸显了以印刷或电子书格式进行自出版的最新发展趋势。与 2015 年相比，2016 年印刷格式的自出版持续增长(11%)，较一年前(34%)有所下滑。电子书格式的自出版则略有下降(所有权登记数量下降了 3%)，但与上年相比，下降幅度变小(上年同比下降幅度为 11%)。

Bowker 标识服务总监 Beat Barblan 指出："总的来说，我们认为这些数字意味着自出版业的持续成熟和稳定。报告还指出，自出版业由三家服务提供商主导，合计占去年出版的印刷和电子书籍的 84%以上。"

Barblan 补充说："跟踪这些趋势，比较这些年份的数据，可以深入了解这一行业。因而，我们能够根据当前的需求，向独立作者提供最好的工具和支持服务。"

(编译自：http://www.bowker.com/news/2017/Self-Publishing-ISBNs-Climbed-8-Between-2015-2016.html)

(本刊讯)